Accepted Manuscript

Data Driven Computing with noisy material data sets

T. Kirchdoerfer, M. Ortiz

PII: S0045-7825(17)30401-2
DOI: http://dx.doi.org/10.1016/j.cma.2017.07.039
Reference: CMA 11558

To appear in: *Comput. Methods Appl. Mech. Engrg.*

Received date : 3 March 2017
Revised date : 17 May 2017
Accepted date : 27 July 2017

Please cite this article as: T. Kirchdoerfer, M. Ortiz, Data Driven Computing with noisy material data sets, *Comput. Methods Appl. Mech. Engrg.* (2017), http://dx.doi.org/10.1016/j.cma.2017.07.039

Data Driven Computing with Noisy Material Data Sets

T. Kirchdoerfer[a], M. Ortiz[a,*]

[a] *Graduate Aerospace Laboratories, California Institute of Technology*
1200 E. California Blvd., Pasadena, Ca, USA, 91125

Abstract

We formulate a Data Driven Computing paradigm, termed max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing and is robust with respect to outliers. Robustness is achieved by means of clustering analysis. Specifically, we assign data points a variable relevance depending on distance to the solution and on maximum-entropy estimation. The resulting scheme consists of the minimization of a suitably-defined free energy over phase space subject to compatibility and equilibrium constraints. Distance-minimizing Data Driven schemes are recovered in the limit of zero temperature. We present selected numerical tests that establish the convergence properties of the max-ent Data Driven solvers and solutions.

Keywords: data science, big data, approximation theory, scientific computing

1. Introduction

Despite the phenomenal growth of scientific computing over the past 50 years, several stubborn challenges have remained foci of extensive research to this day. One of those challenges is *material modelling*. The prevailing and classical scientific computing paradigm has been to calibrate empirical material models using observational data and then use the calibrated material models in calculations. This process of modelling inevitably adds error and uncertainty to the solutions, especially in systems with high-dimensional phase spaces and complex material behavior. This modelling error and uncertainty arises mainly from imperfect knowledge of the functional form of the material laws, the phase space in which they are defined, and from scatter and noise in the experimental data. Simultaneously, advances in experimental science over the past few decades have changed radically the nature of science and engineering from *data-starved* fields

*1200 E. California Blvd., MC 105-50, Pasadena, Ca, USA, 91125
Email address: ortiz@caltech.edu (M. Ortiz)

to, increasingly, *data-rich* fields, thus opening the way for the application of the emerging field of *Data Science* to science and engineering. Data Science currently influences primarily non-STEM fields such as marketing, advertising, finance, social sciences, security, policy, and medical informatics, among others. By contrast, the full potential of Data Science as it relates to science and engineering has yet to be explored and realized.

The present work is concerned with the development of a Data Science paradigm, to be referred to as *Data Driven Computing*, tailored to scientific computing and analysis, cf. [1]. Data Driven Computing aims to formulate initial-boundary-value problems, and corresponding calculations thereof, directly from material data, thus bypassing the empirical material modelling step of traditional science and engineering altogether. In this manner, material modelling empiricism, error and uncertainty are eliminated entirely and no loss of experimental information is incurred. Here, we extend earlier work on Data Driven Computing [1] to random material data sets with finite probability of *outliers*. We recall that the Data Driven Computing paradigm formulated in [1], or *distance-minimizing* Data Driven Computing, consists of identifying as the best possible solution the point in the material data set that is closest to satisfying the field equations of the problem. Equivalently, the distance-minimizing Data Driven solution can be identified with the point in phase space that satisfies the field equations and is closest to the material data set. It can be shown [1] that distance-minimizing Data Driven solutions converge with respect to uniform convergence of the material set. However, distance-minimizing Data Driven solutions can be dominated by *outliers* in cases in which the material data set does not converge uniformly. Distance-minimizing Data Driven solvers are sensitive to outliers because they accord overwhelming influence to the point in the material data set that is closest to satisfying the field equations, regardless of any clustering of the material data points.

The central objective of the present work is to develop a new Data Driven Computing paradigm, to be called max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing and is robust with respect to outliers. Robustness is achieved by means of clustering analysis. Specifically, we assign data points a variable relevance depending on distance to the solution and through maximum-entropy estimation. The resulting scheme consists of the minimization of a free energy over phase space subject to compatibility and equilibrium constraints. We note that this problem is of non-standard type, in that the relevant free energy is a function of state defined over phase space, i. e., a joint function of the driving forces and fluxes of the system. Max-ent Data Driven solutions are robust with respect to outliers because a cluster of data points can override an outlying data point even if the latter is closer to the constraint set that any point in the cluster. The distance-minimizing Data Driven schemes [1] are recovered in the limit of zero temperature. We also develop a simulated annealing scheme that, through an appropriate annealing schedule zeros in on the most relevant data cluster and the attendant solution. We assess the convergence properties of max-ent Data Driven solutions and simulated

annealing solver by means of numerical testing.

The paper is organized as follows. In Section 2, we begin by laying out the connection between Data Science and Scientific Computing that provides the conceptual basis for Data Driven Computing. In Section 3, we turn attention to random material data sets that may contain outliers, or points far removed from the general clustering of the material data points, with finite probability and develop max-ent Data Driven solvers by an appeal to Information Theory and maximum-entropy estimation. In Section 4, we develop a simulated annealing solver that zeros in on the solution, which minimizes a suitably-defined free energy over phase space by progressive quenching. In Section 5, we present numerical tests that assess the convergence properties of max-ent Data Driven solutions with respect to uniform convergence of the material data set. We also demonstrate the performance of Data Driven Computing when the material behavior itself is random, i. e., defined by a probability density over phase space. Finally, concluding remarks and opportunities for further development of the Data Driven paradigm are presented in Section 6.

2. The Data Driven Science paradigm

In order to understand the hooks by which Data Science may attach itself to Scientific Computing, it helps to review the structure of a typical scientific calculation. Of special import to the present discussion is the fundamentally different roles that conservation and material laws play in defining that structure, with the former setting forth hard universal or material-independent constraints on the states attainable by the system and the latter bringing in material specificity open to empirical determination and sampling.

2.1. The 'anatomy' of boundary-value problems

We begin by noting that the field theories that provide the basis for scientific computing have a common general structure. Perhaps the simplest field theory is potential theory, which arises in the context of Newtonian mechanics, hydrodynamics, electrostatics, diffusion, and other fields of application. In this case, the field φ that describes the global state of the system is scalar. The *localization law* that extracts from φ the *local state* at a given material point is $\epsilon = \nabla\varphi$, i. e., the localization operator is simply the gradient of the field, together with essential boundary conditions of the Dirichlet type. The corresponding conjugate variable is the *flux* σ. The flux satisfies the *conservation equation* $\nabla \cdot \sigma = \rho$, where $\nabla\cdot$ is the divergence operator and ρ is a source density, together with natural boundary conditions of the Neumann type. The pair $z = (\epsilon, \sigma)$ describes the local state of the system at a given material point and takes values in the product space $Z = \mathbb{R}^n \times \mathbb{R}^n$, or *phase space*. We note that the phase space, localization and conservation laws are universal, i. e., material independent. We may thus define a material-independent *constraint set* C to be

the set of local states $z = (\epsilon, \sigma)$ consistent with the localization and conservation laws, including corresponding essential and natural boundary conditions.

The localization and conservation laws are closed by appending an appropriate material law. In general, material laws express a relation between fluxes and corresponding *driving forces*. In field theories, the assumption is that local states supply the forces driving the fluxes, leading to material laws of the form $\sigma(\epsilon)$. Often, such material laws are only known imperfectly through a material data set E in phase space Z that collects the totality of our empirical knowledge of the material. Thus, suppose that, in contrast to the classical formulation of initial-boundary-value problems in science and engineering, the material law is imperfectly characterized by a material data point set E. A typical material data set then consists of a finite number of local states, $E = ((\epsilon_i, \sigma_i),\ i = 1, \ldots, N)$. Evidently, for a material data set of this type, the intersection $E \cap C$ is likely to be empty, i. e., there may be no points in the material data set that are compatible with the localization and conservation laws, even in cases when solutions could reasonably be expected to exist. It is, therefore, necessary to replace the overly-rigid characterization of the solution set $S = E \cap C$ by a suitable relaxation thereof.

2.2. Distance-minimizing Data Driven schemes

One such relaxed formulation of Data Driven Computing [1] consists of accepting as the best possible solution the point $z_i = (\epsilon_i, \sigma_i)$ in the material data set E that is closest to the constrained set C, i. e., the point that is closest to satisfying the localization and conservation laws. Closeness is understood in terms of some appropriate distance d defined in phase space Z. The corresponding distance from a local state z to the material data set E is, then: $d(z, E) = \min_{y \in E} d(z, y)$, and the optimal solution is the solution of the minimum problem: $\min_{z \in C} d(z, E)$. Evidently, the data driven problem can also be directly formulated as the double minimization problem: $\min_{z \in C} \min_{y \in E} d(z, y)$. Inverting the order of minimization, we obtain the equivalent data driven problem: $\min_{y \in E} \min_{z \in C} d(z, y)$, or: $\min_{y \in E} d(y, C)$. This reformulation identifies the data driven solution as the point y in the constraint set C that is closest to the material data set E.

2.3. An elementary example

The distance-minimizing Data Driven Computing paradigm just outlined is illustrated in Fig. 1 by means of the elementary example of a elastic bar deforming uniformly under the action of a well-calibrated loading device. In this example, phase space Z is the (ϵ, σ)-plane, the material data set is a point set E in phase space and the constraint set is a straight line C of slope and location determined by the stiffness k of the loading device and the applied displacement u_0. In general, the constraint set C and the material data set E may have empty intersection. However, the distance-minimizing Data Driven solution is

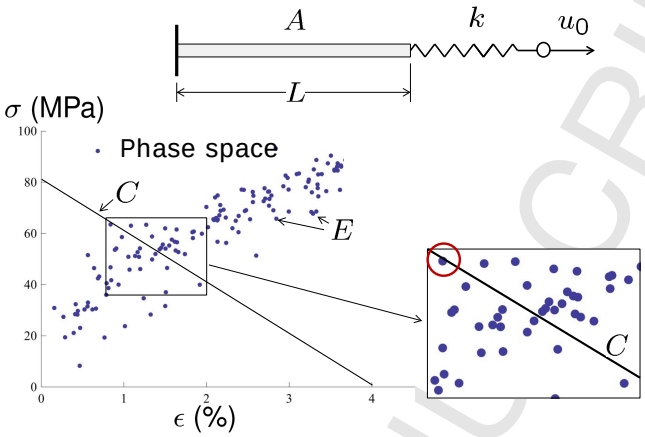


Figure 1: Bar loaded by soft device. The data driven solution is the point in the material data set (circled in red) that is closest to the constraint set.

well defined, though not necessarily uniquely, as the point of the material data set that is closest to the constraint set, circled in red in Fig. 1. The fundamental property of the distance-minimizing Data Driven scheme thus defined is that it makes direct use of the material data set in calculations, entirely bypassing the intermediate modelling step of conventional material identification. In the simple example of the bar loaded by a soft device, it is clear that the distance-minimizing Data Driven solution converges to a classical solution if the data traces a graph in phase space with increasing sampling density, which is a *sanity-check* requirement.

2.4. Uniform convergence

The variational structure of distance-minimizing Data Driven problems confers additional robustness to the solvers and renders them amenable to analysis. By exploiting this connection, it can be shown [1] that distance-minimizing Data Driven solutions converge to classical solutions when the data set approximates a limiting graph in phase space with increasing fidelity, a test case that provides a *sanity check*. Specifically, suppose that the limiting material law is represented by a graph E in phase space, and that a sequence (E_k) of material data sets is such that: i) there is a sequence $\rho_k \downarrow 0$ such that $\mathrm{dist}(z, E_k) \leq \rho_k$, for all $z \in E$, and ii) there is a sequence $t_k \downarrow 0$ such that $\mathrm{dist}(z_k, E) \leq t_k$, for all $z_k \in E_k$. Then, with an additional transversality assumption between the data and the constraint set, it follows [1] that the corresponding sequence (z_k) of distance-minimizing Data Driven solutions converges to the solution z of the classical problem defined by the classical material law E. In addition, if the discrete problem is the result of spatial discretization, e. g., by means of the finite element method, then, under the same assumptions, the sequence (u_k) of

solutions corresponding to a sequence (E_k) of material data sets converges in norm to the classical solution u of the boundary value problem defined by the classical material law E.

3. Probabilistic Data Driven schemes

In practice, material data sets may be random by virtue of inherent stochasticity of the material behavior, experimental scatter inherent to the method of measurement, specimen variability and other factors. Under these conditions, the material data set may contain outliers, or points far removed from the general clustering of the material data points, with finite probability. If one of these outliers happens to be close to the constraint set, it may unduly dominate the distance-minimizing Data Driven solution described in the foregoing. Thus, such solvers, while well-behaved in applications with material data sets with uniformly bounded scatter, may not be sufficiently robust with respect to persistent outliers in other applications. This limitation of distance-minimizing Data Driven solvers points to the need to investigate problems with random material data sets from a probabilistic perspective, with a view to ranking the data points by relevance and importance and understanding the probability distribution of outcomes of interest.

3.1. Data clustering

Distance-minimizing Data Driven solvers are sensitive to outliers because, for any given test solution z in phase space, they accord overwhelming influence to the nearest point in the material data set, regardless of any clustering of the points. Cluster analysis provides a means of mitigating the influence of individual material data points and building notions of data clustering into the Data Driven solver. Cluster analysis can be based on fundamental concepts of Information Theory such as maximum-entropy estimation [2]. Specifically, we wish to quantify how well a point z in phase space is represented by a point z_i in a material data set $E = (z_1, \dots z_n)$. Equivalently, we wish to quantify the relevance of a point z_i in the material data set to a given point z in phase space. We measure the relevance of points z_i in the material data set by means of weights $p_i \in [0, 1]$ with the property

$$\sum_{i=1}^{n} p_i = 1. \tag{1}$$

We wish the ranking by relevance of the material data points to be unbiased. It is known from Information Theory that the most unbiased distribution of weights is that which maximizes Shannon's information entropy [3–5]

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

with the extension by continuity: $0 \log 0 = 0$. In addition, we wish to accord points distant from z less weight than nearby points, i. e., we wish the cost function

$$U(z, p) = \sum_{i=1}^{n} p_i d^2(z, z_i) \tag{3}$$

to be as small as possible. These competing objectives can be combined in the sense of Pareto optimality. The Pareto optima are solutions of the problem

$$\text{For fixed } z\text{, minimize:} \qquad \beta U(z, p) - H(p) \tag{4}$$

$$\text{subject to:} \qquad p_i \geq 0,\ i = 1, \ldots, n; \quad \sum_{i=1}^{N} p_i = 1, \tag{5}$$

where $\beta \in (0, +\infty)$ is a Pareto weight. The solution to this problem is given by the Bolzmann distribution

$$p_i(z, \beta) = \frac{1}{Z(z, \beta)} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}, \tag{6a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}. \tag{6b}$$

The corresponding max-ent Data Driven solver now consists of minimizing the free energy

$$F(z, \beta) = -\frac{1}{\beta} \log Z(z, \beta), \tag{7}$$

over the constraint set C, i. e.,

$$z \in \operatorname{argmin}\{F(z', \beta),\ z' \in C\}. \tag{8}$$

We note that $\beta^{-1/2}$ represents the width of the Bolzmann distribution (6) in phase space. Thus, points in the data set at a distance to z large compared to $\beta^{-1/2}$ have negligible influence over the solution. Conversely, the solution z is dominated by the local cluster of data points in the $\beta^{-1/2}$-neighborhood of z. In particular, outliers, or points outside that neighborhood, have negligible influence over the solution.

For a compact material point set E in a finite-dimensional phase space, the existence of solutions of problem (8) is ensured by the Weierstrass extreme value theorem. We also note that the distance-minimizing Data Driven scheme [1] is recovered in the limit of $\beta \to +\infty$. By analogy to statistical thermodynamics, max-ent Data Driven Computing may be regarded as a *thermalized* extension of distance-minimizing Data Driven Computing. For finite β, all points in the material data set influence the solution, but their corresponding weights diminish with distance to the solution. In particular, the addition of an outlier that is marginally closer to the constraint set C than a large cluster of material data points does not significantly alter the solution, as desired.

4. Numerical implementation

We recall that the max-ent Data Driven problem of interest is to minimize the free energy $F(z)$ (7) over the constraint set C. The corresponding optimality condition is

$$\frac{\partial F}{\partial z}(z,\beta) \perp C, \tag{9}$$

where $\perp$ denotes orthogonality. Assuming

$$d(z,z') = |z - z'|, \tag{10}$$

with $|\cdot|$ the standard norm in $\mathbb{R}^n$, we compute

$$\frac{\partial F}{\partial z}(z,\beta) = \sum_{i=1}^{n} p_i(z,\beta)(z - z_i) = z - \sum_{i=1}^{n} p_i(z,\beta) z_i. \tag{11}$$

Inserting this identity into (9), we obtain

$$z - \sum_{i=1}^{n} p_i(z,\beta) z_i \perp C, \tag{12}$$

which holds if and only if

$$z = P_C\left(\sum_{i=1}^{n} p_i(z,\beta) z_i\right), \tag{13}$$

where P_C is the closest-point projection to C. For instance, if $C = \{f(z) = 0\}$ for some constraint function $f(z)$, (9) may be expressed as

$$\frac{\partial F}{\partial z}(z,\beta) = \lambda \frac{\partial f}{\partial z}(z), \tag{14a}$$

$$f(z) = 0, \tag{14b}$$

where λ is a Lagrange multiplier.

The essential difficulty inherent to problem (9), or (14), is that, in general, the free energy function $F(\cdot,\beta)$ is strongly non-convex, possessing multiple wells centered at the data points in the material data set. Under these conditions, iterative solvers may fail to converge or may return a local minimizer, instead of the global minimizer of interest.

We overcome these difficulties by recourse to *simulated annealing* [6]. The key observation is that the free energy $F(\cdot,\beta)$ is convex for sufficiently small β. Indeed, a straightforward calculation using (10) gives the Hessian matrix as

$$\begin{aligned}\frac{\partial^2 F}{\partial z \partial z}(z) = I &- \beta \sum_{i=1}^{n} p_i(z)(z - z_i) \otimes (z - z_i) \\ &+ \beta \left(\sum_{i=1}^{n} p_i(z)(z - z_i)\right) \otimes \left(\sum_{j=1}^{n} p_j(z)(z - z_j)\right)\end{aligned} \tag{15}$$

Evidently, in the limit of $\beta \to 0$ the Hessian reduces to the identity and the free energy is convex. Indeed, it follows from (6b) that, for $\beta \to 0$,

$$F(z,\beta) - \frac{1}{\beta}\log\frac{1}{n} \sim \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}d^2(z,z_i). \tag{16}$$

The main idea behind simulated annealing is, therefore, to initially set β sufficiently small that $F(\cdot,\beta)$ is convex and subsequently increase it according to some appropriate annealing schedule, with a view to guiding the solver towards the absolute minimizer.

4.1. Fixed-point iteration

We begin by noting that, for fixed β, eq. (13) conveniently defines the following fixed-point iteration,

$$z^{(k+1)} = P_C\left(\sum_{i=1}^{n} p_i(z^{(k)},\beta)z_i\right). \tag{17}$$

We recall that fixed-point iterations $z \leftarrow f(z)$ converge if the mapping $f(z)$ is contractive. Since P_C is an orthogonal projection, it is contractive if the constraint set C is convex, which we assume henceforth. Under this assumption, the mapping (17) is contractive if and only if the mapping

$$z \mapsto \sum_{i=1}^{n} p_i(z,\beta)z_i \equiv g(z,\beta) = z - \frac{\partial F}{\partial z}(z,\beta) \tag{18}$$

is contractive.

Conditions ensuring local contractivity of the mapping $g(\cdot,\beta)$ are given by the following theorem.

Theorem 1. *Suppose that*

$$\frac{1}{\beta} < \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2, \tag{19}$$

where

$$\bar{z} = \sum_{i=1}^{n} p_i(z,\beta)z_i. \tag{20}$$

Then, $g(\cdot,\beta)$ is contractive in a neighborhood of z.

Proof. From the definition (18) of $g(z,\beta)$, we have, after a trite calculation,

$$\frac{\partial g}{\partial z}(z) = \beta\sum_{i=1}^{n} p_i(z,\beta)(z_i - \bar{z}) \otimes (z_i - \bar{z}). \tag{21}$$

Let $u \in Z$, $|u| = 1$. Then,

$$u^T \frac{\partial g}{\partial z}(z) u = \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z_i - \bar{z}) \cdot u\big)^2 \leq \beta \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2. \tag{22}$$

Therefore, by the implicit function theorem, contractivity in a neighborhood of z follows if

$$\beta \sum_{i=1}^{n} p_i(z,\beta)|z_i - \bar{z}|^2 < 1, \tag{23}$$

or, equivalently, if (19) holds. □

Conditions ensuring global contractivity of the mapping $g(\cdot,\beta)$ are given by the following theorem.

Theorem 2. *Suppose that*

$$\frac{1}{\beta} > \frac{1}{n} \sum_{i=1}^{n} |z - z_i|^2, \tag{24}$$

for all $z \in \Omega \subset Z$. Then, $F(\cdot,\beta)$ is convex in Ω.

Proof. Fix $z \in \Omega$ and $\beta > 0$ and let $u \in Z$ be an arbitrary unit vector in phase space. We have

$$u^T \frac{\partial^2 F}{\partial z \partial z} u = 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i) \cdot u\big)^2 + \beta \left(\sum_{i=1}^{n} p_i(z,\beta)(z - z_i) \cdot u \right)^2, \tag{25}$$

which gives the lower bound

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i) \cdot u\big)^2. \tag{26}$$

Maximizing the bound with respect to u, we have

$$\sum_{i=1}^{n} p_i(z,\beta)\big((z - z_i) \cdot u\big)^2 \leq \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2, \tag{27}$$

and, hence,

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \beta \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2. \tag{28}$$

From the max-ent optimality of $p_i(z,\beta)$, we additionally have

$$\frac{\beta}{2} \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 + \sum_{i=1}^{n} p_i(z,\beta) \log p_i(z,\beta) \leq \frac{\beta}{2} \sum_{i=1}^{n} p_i'|z - z_i|^2 + \sum_{i=1}^{n} p_i' \log p_i', \tag{29}$$

for all (p_i') such that

$$p_i' \geq 0, \quad \sum_{i=1}^{n} p_i' = 1. \tag{30}$$

Testing with $p_i' = 1/n$, we obtain

$$\frac{\beta}{2} \sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 + \sum_{i=1}^{n} p_i(z,\beta) \log p_i(z,\beta) \leq \frac{\beta}{2} \frac{1}{n} \sum_{i=1}^{n} |z - z_i|^2 + \log \frac{1}{n} \tag{31}$$

But, by Jensen's inequality,

$$\log \frac{1}{n} \leq \sum_{i=1}^{n} p_i(z,\beta) \log p_i(z,\beta), \tag{32}$$

which, in conjunction with (31) gives

$$\sum_{i=1}^{n} p_i(z,\beta)|z - z_i|^2 \leq \frac{1}{n} \sum_{i=1}^{n} |z - z_i|^2. \tag{33}$$

Inserting this estimate in (28) gives

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 1 - \frac{\beta}{n} \sum_{i=1}^{n} |z - z_i|^2. \tag{34}$$

From this inequality we conclude that

$$u^T \frac{\partial^2 F}{\partial z \partial z} u \geq 0 \tag{35}$$

for all unit vectors u in phase space if

$$1 - \frac{\beta}{n} \sum_{i=1}^{n} |z - z_i|^2 \geq 0, \tag{36}$$

or, equivalently, if inequality (24) is satisfied. □

4.2. Simulated annealing

The general idea of simulated annealing is to evolve the reciprocal temperature jointly with the fixed point iteration according to an appropriate annealing schedule, i. e., we modify (17) to

$$z^{(k+1)} = P_C \left(\sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)}) z_i \right). \tag{37}$$

An effective annealing schedule is obtained by selecting $\beta^{(k+1)}$ so as to ensure local contractivity of the fixed-point mapping. An appeal to Theorem 1 suggests schedules such that

$$\frac{1}{\beta^{(k+1)}} < \sum_{i=1}^{n} p_i(z^{(k)}, \beta^{(k)})|z_i - \bar{z}^{(k)}|^2. \tag{38}$$

By Theorem 1, this choice ensures local contractivity and, hence, convergence of the fixed-point iteration (17). The initial reciprocal temperature $\beta^{(0)}$ may be chosen according to Theorem 2, which ensures that the fixed-point iteration is contractive everywhere.

As already noted, the max-ent Data Driven solution is controlled by its local $\beta^{-1/2}$-neighborhood of points in the data set. Thus, initially the annealing schedule casts a broad net and all points in the data set are allowed to influence the solution. As β grows, that influence is restricted to an increasingly smaller cluster of data points. For large β, the solution is controlled by the points in a certain local neighborhood of the data set determined by the annealing iteration. In particular, the influence of outliers in the data set is eliminated.

5. Numerical tests

We test the properties of max-ent Data Driven Computing by means of the simple example of truss structures. Trusses are assemblies of articulated bars that deform in uniaxial tension or compression. Thus, conveniently, in a truss the material behavior of a bar e is characterized by a simple relation between the uniaxial strain ε_e and uniaxial stress σ_e in the bar. We refer to the space of pairs $z_e = (\varepsilon_e, \sigma_e)$ as the *phase space* of bar e. The state $z = (z_e)_{e=1}^m$, where m is the number of bars in the truss, is subject to the compatibility and equilibrium constraints

$$\epsilon_e = B_e u, \tag{39a}$$

$$\sum_{e=1}^{m} B_e^T w_e \sigma_e = f, \tag{39b}$$

where u is the array of nodal displacements, f is the array of applied nodal forces, the matrices $(B_e)_{e=1}^m$ encode the geometry and connectivity of the truss members and w_e is the volume of member e.

We may metrize the local phase spaces of each member of the truss by means of Euclidean distances derived from the norms

$$|z_e|_e = \left(\mathbb{C}\epsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\right)^{1/2}, \tag{40}$$

for some positive constant $\mathbb{C}$. We may then metrize the global state of the truss by means of the global norm

$$|z| = \Big(\sum_{e=1}^{m} w_e |z_e|_e^2\Big)^{1/2} = \left(\sum_{e=1}^{m} w_e \Big(\mathbb{C}\epsilon_e^2 + \mathbb{C}^{-1}\sigma_e^2\Big)\right)^{1/2} \tag{41}$$

and the associated distance (10). For a truss structure, the point in C closest to a given point z^* in phase space follows from the stationarity condition

$$\delta\left\{\sum_{e=1}^{m} w_e \left(\frac{\mathbb{C}}{2}(B_e u - \epsilon_e^*)^2 + \frac{\mathbb{C}^{-1}}{2}(\sigma_e - \sigma_e^*)^2\right) + \Big(f - \sum_{e=1}^{m} w_e B_e^T \sigma_e\Big)^T \lambda\right\} = 0, \tag{42}$$

where λ is an array of Lagrange multiplier enforcing the equilibrium constraints. The corresponding Euler-Lagrange equations are

$$\sum_{e=1}^{m} w_e B_e^T \mathbb{C}(B_e u - \epsilon_e^*) = 0, \tag{43a}$$

$$\mathbb{C}^{-1}(\sigma_e - \sigma_e^*) - B_e \lambda = 0, \tag{43b}$$

$$\sum_{e=1}^{m} w_e B_e^T \sigma_e = f, \tag{43c}$$

or

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) u = \sum_{e=1}^{m} w_e B_e^T \mathbb{C}\epsilon_e^*, \tag{44a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) \lambda = f - \sum_{e=1}^{m} w_e B_e^T \sigma_e^*, \tag{44b}$$

which define two standard truss equilibrium problems for the linear reference material of modulus $\mathbb{C}$.

We assume that the behavior of each bar e is characterized by a local material data set $E_e = \{z_{i_e} = (\epsilon_{i_e}, \sigma_{i_e}) \in \mathbb{R}^2,\ i_e = 1, \ldots, n_e\}$, where n_e is the number of data points in E_e. For instance, each point in the data set may correspond, e. g., to an experimental measurement. A typical data set is notionally depicted in Fig. 1. The global data set is then the Cartesian product

$$E = E_1 \times \cdots \times E_m. \tag{45}$$

A typical point in such a data set is most convenient indexed as $z_{i_1 \ldots i_m}$, with $i_e = 1, \ldots, n_e$, $e = 1, \ldots, m$, instead of using a single index as in Section 3. The partition function (6b) then takes the form

$$Z(z, \beta) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \mathrm{e}^{-(\beta/2)\sum_{e=1}^{m} d^2(z_e, z_{i_e})}, \tag{46}$$

where the local distance is given by (40). Rearranging terms, (46) may be rewritten in the form

$$\begin{aligned} Z(z,\beta) &= \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} \left(\prod_{e=1}^{m} \mathrm{e}^{-(\beta/2)d^2(z_e,z_{i_e})} \right) \\ &= \prod_{e=1}^{m} \left(\sum_{i_e=1}^{n_e} \mathrm{e}^{-(\beta/2)d^2(z_e,z_{i_e})} \right) \equiv \prod_{e=1}^{m} Z_e(z_e,\beta), \end{aligned} \tag{47}$$

and the total free energy evaluates to

$$F(z,\beta) = -\frac{1}{\beta}\log Z(z,\beta) = \sum_{e=1}^{m} \left(-\frac{1}{\beta}\log Z_e(z_e,\beta) \right) \equiv \sum_{e=1}^{m} F_e(z_e,\beta). \tag{48}$$

We note that the total free energy is additive with respect to the free energies $F_e(z_e,\beta)$ of the members. Finally, the Bolzmann distribution (6) becomes

$$\begin{aligned} p_{i_1,\ldots,i_m}(z,\beta) &= \frac{1}{Z(z,\beta)} \mathrm{e}^{-(\beta/2)\sum_{e=1}^{m} d^2(z_e,z_{i_e})} \\ &= \prod_{e=1}^{m} \left(\frac{1}{Z_e(z_e,\beta)} \mathrm{e}^{-(\beta/2)d^2(z_e,z_{i_e})} \right) \equiv \prod_{e=1}^{m} p_{i_e}(z_e,\beta). \end{aligned} \tag{49}$$

As expected, the local memberwise probability distributions are independent. We also note that the system is assumed to be in thermal equilibrium, i. e., the members are all assumed to be at the same temperature.

In the case of independent local material data sets, eq. (45), the bound (19) specializes to

$$\begin{aligned} \frac{1}{\beta} &< \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} p_{i_1,\ldots,i_m}(z,\beta) \left(\sum_{e=1}^{m} d^2(\bar{z}_e, z_{i_e}) \right) \\ &= \sum_{e=1}^{m} \left(\sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} p_{i_1,\ldots,i_m}(z,\beta) d^2(\bar{z}_e, z_{i_e}) \right) \\ &= \sum_{e=1}^{m} \left(\sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta) d^2(\bar{z}_e, z_{i_e}) \right). \end{aligned} \tag{50}$$

We can exploit this special structure and refine the bound by applying it at the local level, i. e., by requiring

$$\frac{1}{\beta_e} < \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta_e) d^2(\bar{z}_e, z_{i_e}), \tag{51}$$

$e = 1,\ldots,m$, where $1/\beta_e$ represent local temperatures. We can further define an annealing schedule by taking (51) as the basis for local temperature updates

$$\frac{1}{\beta_e^{(k+1)}} = \sum_{i_e=1}^{n_e} p_{i_e}(z_e,\beta^{(k)}) d^2(\bar{z}_e^{(k)}, z_{i_e}), \tag{52}$$

with thermal equilibrium subsequently restored by setting the global temperature to

$$\frac{1}{\beta^{(k+1)}} = \sum_{e=1}^{m} \frac{w_e^{(k+1)}}{\beta_e^{(k+1)}}, \tag{53}$$

with appropriate weights $w_e^{(k+1)}$. In calculations, we specifically choose

$$w_e^{(k+1)} = \frac{\mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}}{\sum_{e=1}^{m} \mathrm{e}^{-\beta_e^{(k)} F_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}} = \frac{Z_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}{\sum_{e=1}^{m} Z_e(\bar{z}_e^{(k)}, \beta_e^{(k)})}. \tag{54}$$

Finally, the initial estimate (24) corresponds to setting

$$p_{i_e}(z_e, \beta) = \frac{1}{n_e}, \tag{55}$$

whereupon (38) becomes

$$\frac{1}{\beta^{(0)}} = \sum_{e=1}^{m} \frac{1}{n_e} \left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e}) \right). \tag{56}$$

As a further control on the annealing rate we set

$$\beta^{(k+1)} = \lambda \tilde{\beta}^{(k+1)} + (1 - \lambda)\beta^{(k)}, \tag{57}$$

where $\tilde{\beta}^{(k+1)}$ is the result of applying the update (53) and λ is an adjustable factor.

A complete list of relations specialized for the case of independent local material data sets is given in Algorithm 1.

Alternative strategies for starting and accelerating simulated-annealing iterations are briefly noted in Section 6.5, but a detailed investigation of such alternatives is beyond the scope of this paper (cf., e. g., [7] for a general discussion of simulated-annealing strategies). The iterative solver operates in two distinct phases. The first phase executes the annealing schedule until the values for β become large. Subsequent to this initial phase, the algorithm proceeds by distance minimization, as in [1], until convergence is achieved.

Algorithm 1 Data-driven solver

Require: Local data sets $E_e = \{z_{i_e},\, i_e = 1,\dots,n_e\}$, B-matrices $\{B_e,\, e = 1,\dots,m\}$, force vector f, parameter λ.

1) Initialize data iteration. Set $k = 0$, compute

$$\bar{z}_e^{(0)} = z_e^{(0)} = \frac{1}{n_e}\sum_{i_e=1}^{n_e} z_{i_e}, \quad \frac{1}{\beta^{(0)}} = \sum_{e=1}^{m}\frac{1}{n_e}\left(\sum_{i_e=1}^{n_e} d^2(\bar{z}_e^{(0)}, z_{i_e})\right). \tag{58}$$

2) Calculate data associations and precalculate for convexity estimate:
for all $e = 1,\dots,m$ **do**
2.1) Set $c_{i_e}^{(k)} = \exp\big(-\beta^{(k)} d^2(z_e^{(k)}, z_{i_e})\big),\ i_e = 1,\dots,n_e$.
2.2) Set $Z_e^{(k)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(k)}$.
2.3) Set $p_{i_e}^{(k)} = c_{i_e}^{(k)}/Z_e^{(k)},\ i_e = 1,\dots,n_e$.
2.4) Set $\bar{z}_e^{(k)} = \sum_{i_e=1}^{n_e} p_{i_e}^{(k)} z_{i_e}$.
2.5) Set $D_e^{(k)} = \sum_{i_e=1}^{n_e} c_{i_e}^{(k)} d^2(\bar{z}_e^{(k)}, z_{i_e})$
end for
3) Solve for $u^{(k+1)}$ and $\eta^{(k+1)}$:

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) u^{(k+1)} = \sum_{e=1}^{m} w_e B_e^T \mathbb{C}\bar{\epsilon}_e^{(k)}, \tag{59a}$$

$$\left(\sum_{e=1}^{m} w_e B_e^T \mathbb{C} B_e\right) \eta^{(k+1)} = f - \sum_{e=1}^{m} w_e B_e^T \bar{\sigma}_e^{(k)}. \tag{59b}$$

4) Progress Schedule:
4.1) Set

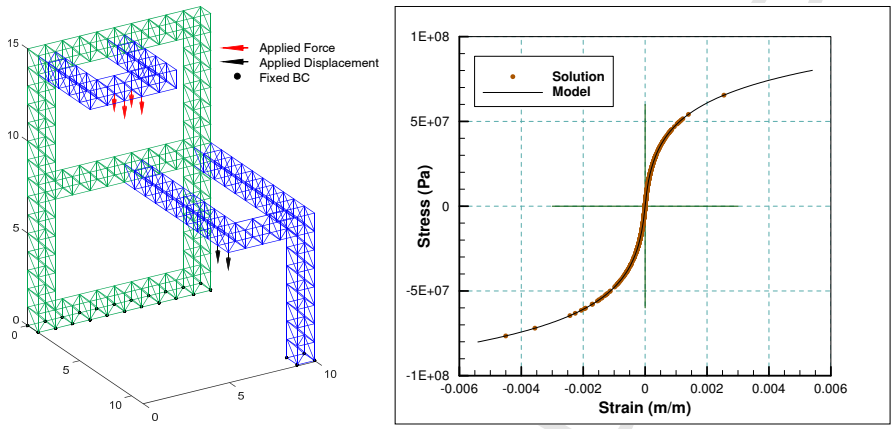
$$\tilde{\beta}^{(k+1)} = \left(\frac{\sum_{e=1}^{m} D_e^{(k)}}{\sum_{e=1}^{m} Z_e^{(k)}}\right)^{-1}. \tag{60}$$

4.2) Set $\beta^{(k+1)} = (1-\lambda)\beta^{(k)} + \lambda\tilde{\beta}^{(k+1)}$.

5) Compute local states $z_{e,k}$:
for all $e = 1,\dots,m$ **do**

$$\varepsilon_e^{(k+1)} = B_e u^{(k+1)}, \qquad \sigma_e^{(k+1)} = \bar{\sigma}_e^{(k+1)} + \mathbb{C} B_e \eta^{(k+1)} \tag{61}$$

end for
8) Test for convergence and cycle the time or data iteration:
if $\{z_e^{(k+1)} = z_e^{(k)},\ e = 1,\dots,m\}$ **then**
exit
else
$k \leftarrow k+1$,
goto (2).
end if

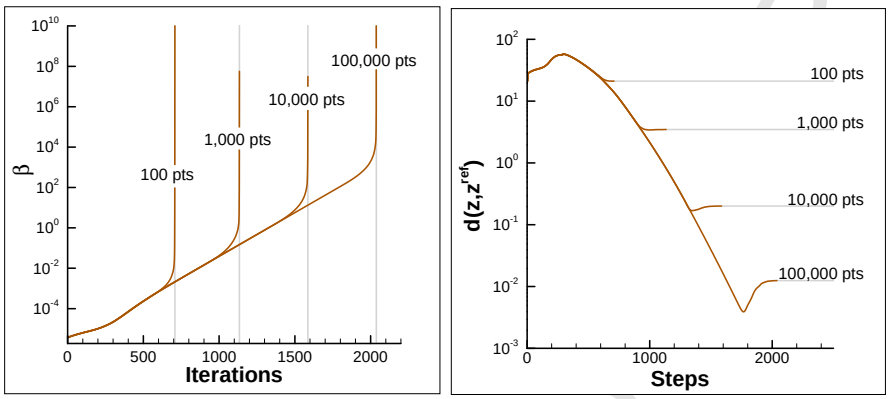


(a) (b)

Figure 2: a) Geometry and boundary conditions of truss test case. b) Base material model with reference solution stress-strain points superimposed.

5.1. Annealing Schedule

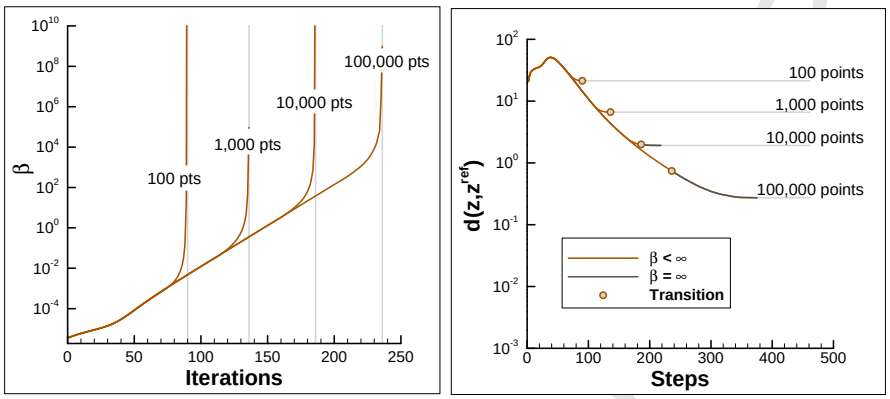
In calculations we consider the specific test case shown in Fig. 2a. The truss contains 1,246 members and is supported and loaded as shown in the figure. By way of reference, we consider the nonlinear stress-strain relation shown in 2b. A Newton-Raphson solution based on that model is readily obtained. The resulting states of all the members of the truss are shown in Fig. 2a superimposed on the stress-strain curve in order to visualize the coverage of phase space entailed by the reference solution.



(a) (b)

Figure 3: Truss test case, , $\lambda = 0.01$. a) Evolution of β through the annealing schedule for different data set sizes. b) Convergence of the max-ent Data Driven solution to the reference solution for the base model depicted in Fig. 2b.

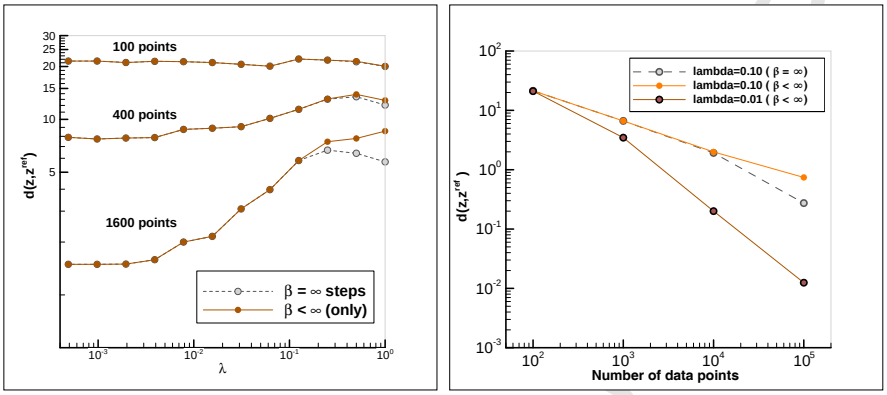
The performance of the solver is shown in Fig. 3. In the present test, data sets are generated by spacing the data points evenly over the strain axis and then evaluating the corresponding stress values from the base model depicted in Fig. 2b. Fig. 3a shows the evolution of β through the annealing schedule for $\lambda = 0.01$. As may be seen from the figure, β grows roughly linearly up to a certain, data set size dependent, number of iterations at which point it diverges rapidly. The stepwise convergence of the simulated annealing iteration is shown in Fig. 3b. As the data set grows in size, the number of iterations to convergence grows correspondingly, as the iteration has to explore a larger data set.



(a) (b)

Figure 4: Truss test case, $\lambda = 0.1$. a) Evolution of β through the annealing schedule for different data set sizes. b) Convergence of the max-ent Data Driven solution to the reference solution for the base model depicted in Fig. 2b.

The influence of the parameter λ on the annealing schedule and the solution is illustrated in Fig. 4, which corresponds to $\lambda = 0.1$. In general, a larger value of λ represents a more aggressive, or faster, annealing schedule, whereas a smaller value represents a more conservative, or slower, annealing schedule. A comparison between Figs. 3 and 4 reveals that, whereas an aggressive annealing schedule indeed speeds up the convergence of the simulated-annealing iteration, it may prematurely freeze the solution around a non-optimal data set cluster, with an attendant loss of accuracy of the solution. Contrariwise, whereas a conservative annealing schedule slows down the convergence of the simulated-annealing iteration, it provides for a more thorough exploration of the data set, resulting in a solution of increased accuracy.



(a) (b)

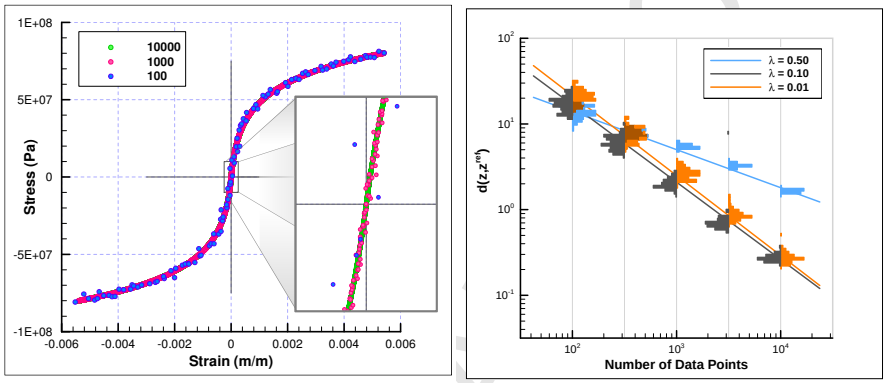
Figure 5: Truss test case. a) Error in the data driven solution relative to the reference solution as a function of λ and data set size. b) Convergence to the reference solution with increasing data set size.

Further evidence of this annealing speed *vs.* accuracy trade-off is collected in Fig. 5. Thus, Fig. 5a shows the error in the data driven solution relative to the reference solution as a function of λ and data set size. As may be seen from the figure, the data driven solution is relatively insensitive to λ for small, or coarse, data sets. This lack of sensitivity owes to the fact that, by virtue of the coarseness of the data set, the simulated-annealing iteration leads to identical, or nearly identical, local data cluster regardless of the value of λ. By contrast, the range of possible limiting local data clusters increases with the size of the data set. Under these conditions, a conservative annealing schedule is more effective at identifying an optimal, or nearly-optimal, local data set cluster, at an attendant improvement in the accuracy of the solution. Fig. 5a also illustrates the beneficial effect of performing a distance-minimizing iteration after quenching (grey symbols) *vs.* stopping the iteration upon quenching (orange symbols). Fig. 5b shows the rates of convergence achieved as a function of λ and the size of the data set. A theorem presented in [1] shows that, for the data sets under consideration, the rate of convergence of distance-minimizing Data Driven solutions with respect to data set size is linear. Fig. 5b suggests that the same rate of convergence is achieved asymptotically by the max-ent Data Driven solutions for sufficiently small λ.

5.2. Uniform convergence of a noisy data set towards a classical material model

Next we consider data sets that, while uniformly convergent to a material curve in phase space, include noise in inverse proportion to the square root of the data set size. To construct a data set consistent with this aim, points are first generated directly from the material curve so that the metric distance between the points is constant. This first sample then has noise added independently

pointwise according to a capped normal distribution in both the strain and stress axes with zero mean and standard deviation in inverse proportion to the square root of the data set size. The resulting data sets converge uniformly to the limiting material curve with increasing number of data points. Fig. 6a illustrates the data sets thus generated when the limiting model is as shown in Fig. 2b.



(a) (b)

Figure 6: Truss test case. a) Random data sets generated according to capped normal distribution centered on the material curve of Fig. 2b with standard deviation in inverse proportion to the square root of the data set size. b) Convergence with respect to data set size of error histograms generated from 100 material set samples.

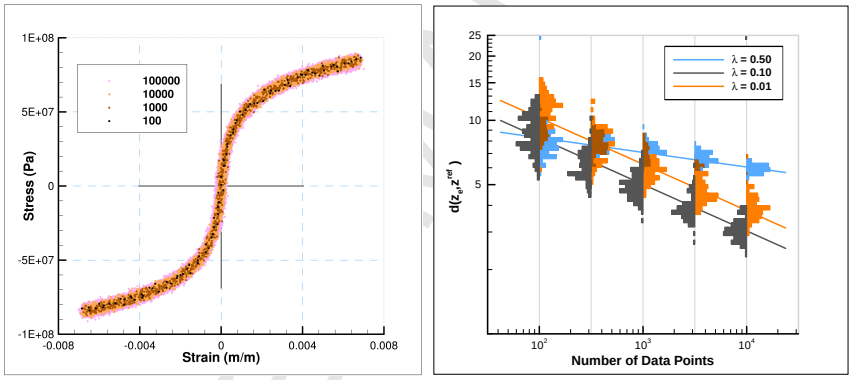
A convergence plot of error *vs.* data set size in shown in Fig. 6b, with error defined as the distance between the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 100 randomly generated data set samples. We again recall that, given the capped structure of the data sets under consideration, distance-minimizing Data Driven solutions converge to the limiting classical solution as $N^{-1/2}$, with N the size of the data set [1]. Surprisingly, an analysis of Fig. 6b suggests that, for sufficiently small λ, the max-ent Data Driven solutions converge as N^{-1} instead, i. e., they exhibit a linear rate of convergence with the data set size.

The random sampling of the data sets also raises questions of convergence in probability. It is interesting to note from Fig. 6b that both the mean error and the standard deviation of the error distribution converge to zero with increasing data set size. As already noted, the mean error exhibits a linear rate of convergence. The roughly constant width of the error histograms in log-log coordinates, suggests that the standard deviation of the error also converges to zero linearly with increasing data set size. These two observations together suggest that the error distribution obtained from a capped normal sampling of a material reference curve converges with sample size to the Dirac distribution centered at zero in both mean and in mean square, hence in probability

[8].

5.3. Random data sets with fixed distribution about a classical material model

A different convergence scenario arises in connection with random material behavior described by a *fixed* probability measure μ in phase space. Specifically, given a set E in phase space, $\mu(E)$ is the probability that a fair test return a state $z \in E$. By virtue of the randomness of the material behavior, the solution becomes itself a random variable. We recall that the constraint set C is the set of states z in phase space that are compatible and in equilibrium. When the material behavior is random and is characterized by a probability measure μ in phase space, the solution must be understood in probabilistic terms and may be identified with the conditional probability $\mu \llcorner C$ of μ conditioned to C. The corresponding question of convergence then concerns whether the distribution of Data Driven solutions obtained by sampling μ by means of data sets of increasing size converges in probability to $\mu \llcorner C$.



(a) (b)

Figure 7: Truss test case. a) Random data sets generated according to normal distribution centered on the material curve of Fig. 2b with constant standard deviation independent of the data set size. b) Convergence with respect to data set size of error histograms generated from 100 material set samples.

While a rigorous treatment of convergence in probability is beyond the scope of this paper, we may nevertheless derive useful insights from numerical tests. We specifically assume that μ is the cartesian product of member-wise measures μ_e characterizing the material behavior of each bar e. Specifically, given a set E_e in the phase space of member e, $\mu_e(E_e)$ is the probability that a fair test of member e return a state $z_e \in E_e$. In accordance with this representation, in calculations we generate data sets member-wise from a zero-mean normal distribution that is no longer capped and whose standard deviation is held constant. Fig. 7a illustrates the data sets thus generated when the base model is as shown in Fig. 2b.

Since the probability measure μ_e is generated by *adding* zero-mean normal random displacements to the base model in phase space, and since the constraint set C is *linear*, the conditional probability $\mu \llcorner C$ is itself centered on the base model. Hence, its mean value $\bar{z}$ necessarily coincides with the classical solution. This property is illustrated in Fig. 7b, which shows a convergence plot of error *vs.* data set size, with error defined as the distance between the max-ent Data Driven solution and the classical solution. For every data set size, the plot depicts histograms of error compiled from 100 randomly generated data set samples. As may be seen from the figure, the mean value of the histograms converges to zero with data set size, which is indicative of convergence in mean of the sampled max-ent Data Driven solutions. The rate of convergence of the mean error is computed to be of the order of 0.22. Interestingly, this rate of convergence is considerably smaller than the linear convergence rate achieved for the capped normal noise distributions considered in the preceding section. The slower rate of convergence may be attributable to the wider spread of the data about its mean, though the precise trade-off between convergence and uncertainty remains to be elucidated rigorously. Finally, we note from Fig. 7b that, as in the case of capped normal noise, an overly fast annealing schedule results in a degradation of the convergence rate.

6. Summary and discussion

We have formulated a Data Driven Computing paradigm, which we have termed max-ent Data Driven Computing, that generalizes distance-minimizing Data Driven Computing of the type proposed in [1] and is robust with respect to outliers. Robustness is achieved by means of clustering analysis. Specifically, we assign data points a variable relevance depending on distance to the solution and through maximum-entropy estimation. The resulting problem consists of the minimization of a suitably-defined free energy over phase space subject to compatibility and equilibrium constraints. The problem is non-standard in the sense that the relevant Data-Driven free energy is defined jointly over driving forces and fluxes. The distance-minimizing Data Driven schemes [1] are recovered in the limit of zero temperature. We have also developed a simulated annealing solver that delivers the solution through a suitably-defined quenching schedule. Finally, we have presented selected numerical tests that establish the good convergence properties of the max-ent Data Driven solutions and solvers.

We conclude by framing Data Driven Computing within the context of past and present efforts to automate the connection between data and material models and expounding on how Data Driven Computing differs from said efforts. We also point out a number of possible enhancements of the approach that define worthwhile directions of further research.

6.1. Irreducibility to classical material laws

As already noted, the Data-Driven free energy (7) and the associated problem (8) are non-standard. Thus, if $z = (\epsilon, \sigma)$, with ϵ the collection of local states of the system and σ the corresponding fluxes, the classical free energy is a function of state of the form $A(\epsilon, \beta)$, i. e., it is a function of the driving forces and temperature, and the fluxes follow as $\sigma = \partial_\epsilon A(\epsilon, \beta)$. The corresponding classical problem consists of minimizing $A(\epsilon, \beta)$ with respect to the driving forces ϵ subject to compatibility constraints. Correspondingly, the classical Gibbs energy is a function of state of the form $G(\sigma, \beta)$, i. e., a function of the fluxes and temperature, and the driving forces follow as $\epsilon = \partial_\sigma G(\sigma, \beta)$. The corresponding classical problem consists of minimizing $G(\sigma, \beta)$ with respect to the fluxes σ subject to equilibrium constraints. By contrast, here the relevant free energy (7) is a function $F(z, \beta)$ defined over the entire phase space, i. e., is a joint function of the driving forces and fluxes. The corresponding Data-Driven problem consists of minimizing $F(z, \beta)$ with respect to z subject to compatibility and equilibrium constraints simultaneously. Of course, it is possible to define an effective free energy through a partial minimization of $F(z, \beta)$ with respect to the fluxes, i. e.,

$$A(\epsilon, \beta) = \min \{F\big((\epsilon, \sigma), \beta\big), (\epsilon, \sigma) \in C\}, \tag{62}$$

with $A = +\infty$ is no minimizer exists. The corresponding effective classical problem then consists of minimizing $A(\epsilon, \beta)$ with respect to the driving forces ϵ. Likewise, it is possible to define an effective Gibbs energy through a partial minimization of $F(z, \beta)$ with respect to the driving forces, i. e.,

$$G(\sigma, \beta) = \min \{F\big((\epsilon, \sigma), \beta\big), (\epsilon, \sigma) \in C\}, \tag{63}$$

with $G = +\infty$ is no minimizer exists. The corresponding effective classical problem then consists of minimizing $G(\sigma, \beta)$ with respect to the fluxes σ. However, we note that these effective energies are *global* and do not correspond to a classical *local* material law in general. For instance, consider a finite-dimensional problem, such as the truss example developed in the foregoing, with compatibility and equilibrium constraints that can be satisfied identically through the representations

$$\epsilon = Bu, \tag{64a}$$

$$\sigma = A\varphi, \tag{64b}$$

where u is a displacement vector, B is a discrete strain operator, φ is an Airy potential vector and A is a discrete Airy operator, with the properties

$$B^T A = 0, \tag{65a}$$

$$A^T B = 0, \tag{65b}$$

which identify B^T and A^T as the discrete equilibrium and compatibility operators, respectively. In this representation, the Data-Driven problem becomes

$$F(Bu, A\varphi, \beta) \to \min! \tag{66}$$

The corresponding Euler-Lagrange equations are

$$A^T \frac{\partial F}{\partial \sigma}(Bu, A\varphi, \beta) = 0, \tag{67a}$$

$$B^T \frac{\partial F}{\partial \epsilon}(Bu, A\varphi, \beta) = 0, \tag{67b}$$

which represent the discrete compatibility and equilibrium equations, respectively. Evidently, it is now possible to eliminate the Airy potential vector φ using the compatibility equations (67a) to define a reduced equilibrium problem in the displacement vector u, or, alternatively, eliminate the displacement vector u using the equilibrium equations (67b) to define a reduced compatibility problem in the Airy potential vector φ. However, these reduced problems remain non-classical in that they are *non-local*, i. e., they do not correspond to any local member-wise material law in general.

6.2. Material Informatics

There has been extensive previous work focusing on the application of Data Science and Analytics to material data sets. The field of Material Informatics (cf., e. g., [9–18]) uses data searching and sorting techniques to survey large material data sets. It also uses machine-learning regression [19, 20] and other techniques to identify patterns and correlations in the data for purposes of combinatorial materials design and selection. These approaches represent an application of standard sorting and statistical methods to material data sets. While efficient at looking up and sifting through large data sets, it is questionable that any real epistemic knowledge is generated by these methods. What is missing in Material Informatics is an explicit acknowledgement of the field equations of physics and their role in constraining and shaping material behavior. By way of contrast, such field equations play a prominent role in the Data Driven Computing paradigm developed in the present work.

6.3. Material identification

There has also been extensive previous work concerned with the use of empirical data for parameter identification in prespecified material models, or for automating the calibration of the models. For instance, the Error-in-Constitutive-Equations (ECE) method is an inverse method for the identification of material parameters such as the Youngs modulus of an elastic material [21–30]. While such approaches are efficient and reliable for their intended application, namely, the identification of material parameters, they differ from Data Driven Computing in that, while material identification schemes aim to determine the parameters of a prespecified material law from experimental data, Data Driven Computing dispenses with material models altogether and uses fundamental material data directly in the formulation of initial-boundary-value problems and attendant calculations thereof.

6.4. Data repositories

A number of repositories are presently in existence aimed at data-basing and disseminating material property data, e. g., [31–34]. However, it is important to note that the existing material data repositories archive parametric data that are specific to prespecified material models. For instance, a number of repositories rely on parameterizations of standard interatomic potentials, such as the embedded-atom method (EAM), and archive data for a wide range of materials systems. Evidently, such data are strongly biased by—and specific to—the assumption of a specific form of the interatomic potential. By way of contrast, Data Driven Computing is based on fundamental, or *model-free*, material data only. Thus, suppose that the problem of interest is linear elasticity. In this case, the field equations are the strain-displacement and the equilibrium relations, and the local states are described by a strain tensor and a corresponding stress tensor. It thus follows that, in this case, model-free fundamental data consists of points in strain-stress space, or *phase space*. By relying solely on fundamental data, Data Driven Computing requires no *a priori* assumptions regarding particular forms, and parameterizations thereof, of material models.

6.5. Implementation Improvements

This paper has focused on a particular definition of the annealing schedule as a means to implementing the new class of max-ent Data Driven solvers. It remains easily within the bounds of expectation that improvements in the schedule definition could lead to reductions in the number of iterations and improvements in annealing convergence rates. A number of other implementation improvements are equally worthy of examination. At present, sums over entire data sets for each material point were calculated without simplification or truncation. However, early stages of the annealing schedule could easily be performed on subsampled or summarized data sets due to the nonlocal nature of the calculations. Late stages of the annealing schedule could easily truncate sums over the data set through the use of cutoff radii pegged to $\beta^{-1/2}$. These and other considerations are likely to play an important role in the progression towards efficient and scalable implementations of the method.

6.6. Data coverage, sampling quality, adaptivity

Data Driven solvers provide, as a by-product, useful information regarding data coverage and sampling quality. Specifically, suppose that z is a Data Driven solution and z_e is the corresponding local state at material point e. Then, the distance $d_e(z_e, E_e)$ supplies a measure of how well the local state z_e is represented within the local material data set E_e. For any given material data set, a certain spread in the values of $d_e(z_e, E_e)$ may be expected, indicating that certain local states in a solution are better sampled than others. Specifically, local states with no nearby data points result in high values of $d_e(z_e, E_e)$, indicative of poor coverage by the material data set. Thus, the analysis of the local values

$d_e(z_e, E_e)$ of the distance function provides a means of improving material data sets adaptively for particular applications. Evidently, the optimal strategy is to target for further sampling the regions of phase space corresponding to the local states with highest values of $d_e(z_e, E_e)$. In particular, local states lying far from the material data set, set targets for further testing. In this manner, the material data set may be adaptively expanded so as to provide the best possible coverage for a particular application. Care would need to be taken in applying such adaptivity in the presence of an annealing schedule so as to prevent the adaptive sampling from biasing the solution.

6.7. Data quality, error bounds, confidence

Not all data are created equal, some data are of higher quality than others. In general, it is important to keep careful record of the pedigree, or ancestry, of each data point and to devise metrics for quantifying the level of confidence that can be placed on the data [35]. The confidence level in a material data point z_i can be quantified by means of a confidence factor $c_i \in [0, 1]$, with $c_i = 0$ denoting no confidence and $c_i = 1$ denoting full confidence. The weighting of the data points can then be modified to

$$p_i(z, \beta) = \frac{c_i}{Z(z, \beta)} \mathrm{e}^{-(\beta/2) d^2(z, z_i)}, \tag{68a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} c_i \mathrm{e}^{-\beta d^2(z, z_i)}, \tag{68b}$$

which effectively factors the confidence factors into the calculations. In addition, material data obtained through experimental measurements often comes with error bounds attached. The standard error of a measurement of mean z_i is normally identified with its standard deviation s_i. In such cases, assuming the distribution of measurements to be Gaussian we obtain the distribution of weights

$$p_i(z, \beta) = \frac{1}{Z(z, \beta)} \mathrm{e}^{-1/2(s_i^2 + 1/2\beta)^{-1} d^2(z, z_i)}, \tag{69a}$$

$$Z(z, \beta) = \sum_{i=1}^{n} \mathrm{e}^{-1/2(s_i^2 + 1/2\beta)^{-1} d^2(z, z_i)}. \tag{69b}$$

Again, this simple device effectively factors the experimental error bounds into the calculations.

Acknowledgements

The support of Caltech's Center of Excellence on High-Rate Deformation Physics of Heterogeneous Materials, AFOSR Award FA9550-12-1-0091, is gratefully acknowledged. We gratefully acknowledge helpful discussions with H. Owhadi and T. J. Sullivan.

Bibliography

[1] T. Kirchdoerfer, M. Ortiz, Data-driven computational mechanics, Computer Methods in Applied Mechanics and Engineering 304 (2016) 81–101.

[2] A. I. Khinchin, Mathematical foundations of information theory, Dover Publications, New York,, 1957.

[3] C. E. Shannon, Communication theory of secrecy systems, Bell System Technical Journal 28 (4) (1949) 656–715.

[4] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (3) (1948) 379–423.

[5] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (4) (1948) 623–656.

[6] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.

[7] P. J. M. v. Laarhoven, E. H. L. Aarts, Simulated annealing : theory and applications, Kluwer Academic Publishers, Boston, MA, 1987.

[8] P. Billingsley, Probability and measure, Wiley, Hoboken, N.J., 2012.

[9] K. Rajan, M. Zaki, K. Bennett, Informatics based design of materials., Abstracts of Papers of the American Chemical Society 221 (2001) U464–U464.

[10] K. Rajan, Materials informatics, Materials Today 8 (10) (2005) 38–45.

[11] S. Broderick, C. Suh, J. Nowers, B. Vogel, S. Mallapragada, B. Narasimhan, K. Rajan, Informatics for combinatorial materials science, Jom 60 (3) (2008) 56–59.

[12] K. Rajan, Materials informatics part I: A diversity of issues, Jom 60 (3) (2008) 50–50.

[13] K. Rajan, Informatics and integrated computational materials engineering: Part II, Jom 61 (1) (2009) 47–47.

[14] K. Rajan, Materials informatics how do we go about harnessing the "big data" paradigm?, Materials Today 15 (11) (2012) 470–470.

[15] K. Rajan, Materials informatics: The materials "gene" and big data, Annual Review of Materials Research, Vol 45 45 (2015) 153–169.

[16] S. Broderick, K. Rajan, Informatics derived materials databases for multifunctional properties, Science and Technology of Advanced Materials 16 (1) (2015) 013501.

[17] S. R. Kalidindi, Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials, International Materials Reviews 60 (3) (2015) 150–168.

[18] S. R. Kalidindi, M. De Graef, Materials data science: Current status and future outlook, Annual Review of Materials Research, Vol 45 45 (2015) 171–193.

[19] C. M. Bishop, Pattern recognition and machine learning, Springer, New York, 2006.

[20] I. Steinwart, A. Christmann, Support vector machines, 1st Edition, Springer, New York, 2008.

[21] S. Guchhait, B. Banerjee, Constitutive error based material parameter estimation procedure for hyperelastic material, Computer Methods in Applied Mechanics and Engineering 297 (2015) 455–475.

[22] B. Banerjee, T. F. Walsh, W. Aquino, M. Bonnet, Large scale parameter estimation problems in frequency-domain elastodynamics using an error in constitutive equation functional, Computer Methods in Applied Mechanics and Engineering 253 (2013) 60–72.

[23] P. Feissel, O. Allix, Modified constitutive relation error identification strategy for transient dynamics with corrupted data: The elastic case, Computer Methods in Applied Mechanics and Engineering 196 (13-16) (2007) 1968–1983.

[24] M. Ben Azzouna, P. Feissel, P. Villon, Robust identification of elastic properties using the modified constitutive relation error, Computer Methods in Applied Mechanics and Engineering 295 (2015) 196–218.

[25] T. Merzouki, H. Nouri, F. Roger, Direct identification of nonlinear damage behavior of composite materials using the constitutive equation gap method, International Journal of Mechanical Sciences 89 (2014) 487–499.

[26] J. E. Warner, M. I. Diaz, W. Aquino, M. Bonnet, Inverse material identification in coupled acoustic-structure interaction using a modified error in constitutive equation functional, Computational Mechanics 54 (3) (2014) 645–659.

[27] M. A. Aguilo, L. Swiler, A. Urbina, An overview of inverse material identification within the frameworks of deterministic and stochastic parameter estimation, International Journal for Uncertainty Quantification 3 (4) (2013) 289–319.

[28] N. Promma, B. Raka, M. Grediac, E. Toussaint, J. B. Le Cam, X. Balandraud, F. Hild, Application of the virtual fields method to mechanical characterization of elastomeric materials, International Journal of Solids and Structures 46 (3-4) (2009) 698–715.

[29] F. Latourte, A. Chrysochoos, S. Pagano, B. Wattrisse, Elastoplastic behavior identification for heterogeneous loadings and materials, Experimental Mechanics 48 (4) (2008) 435–449.

[30] H. M. Nguyen, O. Allix, P. Feissel, A robust identification strategy for rate-dependent models in dynamics, Inverse Problems 24 (6).

[31] The NoMaD Repository, http://nomad-repository.eu/cms/.

[32] The Materials Project, https://materialsproject.org/.

[33] The Knowledgebase of Interatomic Models, https://openkim.org/.

[34] The NIST Materials Genome Initiative, https://mgi.nist.gov/materials-data-repository/.

[35] A. R. Newman, Confidence, pedigree, and security classification for improved data fusion, Proceedings of the Fifth International Conference on Information Fusion, Vol II (2002) 1408–1415.